# PRÁCTICA 1

# ¿CÓMO PODEMOS CAPTURAR LOS DATOS DE LA WEB?

## TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

Autores:

Noel Casado Soler

David Moliner Mateu

# Índice

# 1. Contexto

Partimos de la premisa de ser una empresa que se dedica a la restauración y personalización de motocicletas de tipo Custom. El proyecto de web scraping parte de la necesidad de encontrar los modelos más buscados al mejor precio para, después de reacondicionarlos, ponerlos a la venta.

Para ello, hemos querido investigar la disponibilidad de motos de segunda mano del tipo Custom. Existe multitud de páginas web que se dedican a esto, pero hemos elegido la página web de https://www.moto-ocasion.com tras haber comprobado en el archivo "robots.txt" que esta permitía realizar *web scraping* y a la vez satisfacía nuestras necesidades en cuanto a la información que queremos recolectar.

# 2. Título

El título elegido para este dataset es "Motos en venta de tipo Custom".

# 3. Descripción del dataset

El dataset generado para esta Práctica 1 agrupa todas las motos en venta en el portal de compraventa de motos "moto-ocasion.com" que sean de tipo Custom, tal y como indica el título elegido para este.

# 4. Representación gráfica

## 5. Contenido

Para cada moto, los campos que incluye el dataset son los siguientes:

- **Precio:** precio de venta de la moto.
- **Tipo:** tipo de moto, en este caso siempre deberá ser Custom.
- **Marca:** fabricante de la moto.
- **Modelo:** modelo de la moto.
- **Carnet:** carné requerido para poder conducir la moto.
- **Estado:** estado en el que se encuentra la moto.
- **Kilómetros:** kilómetros que ha realizado la moto.
- **Combustible:** qué tipo de combustible utiliza.
- **Cilindrada:** cilindrada del motor de la moto.
- **Transmisión:** tipo de transmisión del cambio de marchas.
- **Año:** año de fabricación de la moto.
- **Color:** color de la moto.
- **Disponibilidad:** disponibilidad para comprar la moto.

En cuanto al periodo de tiempo de los datos, el dataset se genera en el momento de ejecutar el script, por lo que los datos corresponderán a ese mismo instante no existiendo la posibilidad de tener un dataset histórico.

## 6. Propietario

El propietario del conjunto de datos es la empresa Moto Ocasión Europa S.L., tal y como consta en el apartado de Política de privacidad de su página web (https://www.moto-ocasion.com/politica-de-privacidad/).

En otros análisis de precios de venta de vehículos de segunda mano, aunque no tan dirigidos a un segmento específico como las motos de tipo Custom, la necesidad se centra bien en automatizar la toma de información de vehículos de ocasión en los diferentes portales dos veces al día creando alertas de productos idóneos de acuerdo con los criterios establecidos por el concesionario[1], bien en realizar un análisis de datos para comparar entre los distintos precios que hay en las plataformas web para encontrar la mejor alternativa calidad/precio[2].

No se obtiene en este proceso ningún tipo de dato personal, por lo que no se vulnera la privacidad de ninguna persona.

La página no requiere inicio de sesión ni aceptación de términos y condiciones para acceder a la información, por lo que no se produce ningún incumplimiento en este aspecto.

No hay material protegido por derechos de autor, por lo que no se comete ninguna infracción de estos derechos.

Tras haber revisado el archivo “robots.txt”, se ha comprobado que esta web permitía realizar *web scraping.*

[1] https://datstrats.com/blog/bbdd-coches-segunda-mano/

[2] https://medium.com/saturdays-ai/web-scraping-para-machine-learning-71f7e673bcf3
https://rpubs.com/Juve_Campos/scrapVehiculosMercadoLibre

No existen disposiciones sobre términos de uso, y la información que se rastrea es totalmente pública.

Se utilizan mecanismos para no sobrecargar al servidor con peticiones durante el proceso de carga que puedan influir en el rendimiento del sitio web.

Aunque en principio una empresa podría utilizar la información para fines comerciales, estos fines coincidirían con los originales de la página, por lo que no se causa ningún perjuicio material ni económico.

## 7. Inspiración

Al igual que en el resto de los análisis similares, el dataset resultante es interesante al permitirnos saber de una forma rápida qué motos hay disponibles del tipo que nos interesa y a qué precio en la página web de compraventa.

Sería útil combinar este dataset con datasets obtenidos de otras páginas web de compraventa de motos utilizando otros proyectos de *web scraping*, con tal de generar una base de datos que contenga los datos de varios portales. Podríamos identificar anuncios repetidos y tendríamos una visión más amplia del mercado.

## 8. Licencia

Se escoge la licencia CC BY-NC-SA 4.0 que tiene las siguientes características:

**Atribución**: debe dar el crédito adecuado, proporcionar un enlace a la licencia e indicar si se realizaron cambios. Puede hacerlo de cualquier manera razonable, pero no de ninguna manera que sugiera que el licenciante lo respalda a usted o a su uso.

**No comercial**: no puede utilizar el material con fines comerciales.

**Compartir Igual**: si remezclas, transformas o construyes sobre el material, debes distribuir tus contribuciones bajo la misma licencia que el original.

**Sin restricciones adicionales**: no puede aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros de hacer cualquier cosa que la licencia permita.

Se estima que, dado el carácter pedagógico e investigador del trabajo, es la licencia más adecuada.

## 9. Código

El código se divide en dos archivos:

- **main.py**: script que inicia el *web scraping* y almacena en un archivo .csv el dataset resultante.
- **scraper.py**: script que contiene el código que se encarga de la extracción de los datos de las motos.

Es en este segundo archivo donde definimos la página web por la cual empezaremos a extraer los datos.

En la página inicial tenemos un listado con varias motos, y podemos observar que el resto de los resultados de la búsqueda está paginado.

Primero, entraremos en cada moto de la página y extraeremos los datos que nos interesan; una vez acabemos con las motos que hay en la página en la que nos encontramos, inspeccionaremos si hay página siguiente.

Revisando el código html de la página, vemos que el enlace a la página siguiente es una etiqueta <a> con la clase css “next”. Se busca la etiqueta y se llama recursivamente a la función que obtiene los datos de la página.

Este proceso se repetirá hasta que se acaben las páginas de la búsqueda realizada.

La página web no presenta ninguna dificultad en particular a la hora de recolectar los datos. La única medida que se ha tomado es la de añadir un retardo entre peticiones, para no saturar la página web y evitar ser baneados, y la de utilizar diferentes user-agents para hacer creer a la página web que todas las peticiones no proceden del mismo usuario.

Las librerías específicas utilizadas para este proyecto son las siguientes

(obtenido con **pipreqs**):

- **beautifulsoup4** 4.10.0
- **requests** 2.26.0

y los módulos **os**, **random** y **time** de la biblioteca estándar de Python.

Se puede obtener el código del proyecto en el siguiente repositorio git:

https://github.com/davidmoma/motoscraper

## 10. Dataset

El enlace del DOI del dataset es el siguiente:

https://doi.org/10.5281/zenodo.7335349

## 11. Vídeo

El enlace del vídeo explicativo de la práctica es el siguiente:

https://drive.google.com/file/d/1dqT5Nzkrr-IO2WKSmI4wncAz7gpOAq9J/view?usp=share_link

# 12. Tabla de contribuciones

| Contribuciones | Firma |
|---|---|
| **Investigación previa** | NCS, DMM |
| **Redacción de las respuestas** | NCS, DMM |
| **Desarrollo del Código** | NCS, DMM |
| **Participación en el video** | NCS, DMM |